Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN:2799-1172
Vol: 04, No. 03, April-May 2024
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.43.18.30

Early Warning System of Attrition in the BPO Industry Using Machine Learning Classification Models

Sandrilito Abogada[1*], Laurence Usona[2]

[1,2]Graduate School, Polytechnic University of the Philippines, Philippines.*

Email: [2]lpusona@pup.edu.ph
Corresponding Email: [1]sandy.abogada12@gmail.com*

Received: 28 November 2023 **Accepted:** 14 February 2024 **Published:** 01 April 2024

Abstract: Employee attrition is one of the factors affecting gross margin erosion in the BPO industry, the fastest-growing industry in the Philippines, due to hiring and training costs. The cost of employee attrition depends on the employee's role and salary/wage level. This study proposed shifting the retention approach from reactive to proactive with the use of an early warning system for employee attrition. The early warning system was powered by Machine Learning Classification Models. The data used in this study are employees hired in 2021 and 2022 from one of the Telco/Communication programs in the BPO Industry. The data attributes considered in this study are composed of Employee status, Employee Performance, Employee Satisfaction, Payroll, Time Off History, Schedule, and employee Observation data. The data is trained and tested in the classification models (Decision Trees, rule-based classification, naïve Bayes, KNN, Logistic Regression, and Random Forest). Models are evaluated using the classification performance metrics (AUC, Accuracy, Precision, Recall, and F1 Score). The model with the highest predictive accuracy is selected and deployed to produce employee classification (Risk of Termination, Neutral, and Positive). This study mainly helps the company reduce turnover and costs and increase gross margin with the help of the early warning system that can predict the status of the employees using significant indicators.

Keywords: Employee Turnover, HR Analytics, Machine Learning Models, Attrition Indicators.

1. INTRODUCTION

Employee attrition is one of the factors affecting gross margin erosion in the BPO industry, the fastest-growing industry in the Philippines, due to hiring and training costs. The cost of employee attrition depends on the employee's role and salary or wage level. Based on a study conducted by the Center for American Progress (n.d.) on Employee Turnover Costs", revealed that costs have been set at 16%–213% of their annual salary.

Reactive retention, a conventional strategy employed by employers, involves making efforts to retain an employee only after they have expressed their intent to leave the company [1],[3]. However, the efficacy of these traditional retention methods is notably inadequate, as evidenced by historical company data revealing that approximately 70% of employees who have expressed an intention to depart ultimately do so [1]. Past research further highlights a multitude of factors that contribute to employee attrition, including stress, demographics, performance metrics, leadership quality, job satisfaction, job content, interactions with colleagues, and compensation packages [2],[4],[6],[7],[8],[9].

These drivers are mainly available in the BPO industry and can be used to build an early warning system based on a machine learning classification model to predict whether an employee will leave the company or not based on the indicators of churn. This approach will help the organization prevent employee turnover by identifying which employees are at high risk of leaving and implementing proactive engagement with the identified individuals.
This study aimed to address the following objectives:

- Identify the top causal factors for attrition in the BPO industry (Early Life and Production Employees).
- Understand the effects of the top causal factors and predictors of attrition.
- Identify the best classification model that predicts whether the employee will leave the company in the next month.
- Deploy an early warning system that helps operations identify which employees need retention intervention.

2. RELATED WORK

The Business Process Outsourcing (BPO) industry has rapidly emerged as a significant contributor to economies worldwide. However, the industry's success is often marred by the challenges of employee attrition, prompting researchers to delve into the complex interplay of factors and strategies that influence retention. This review navigates through a constellation of studies that illuminate the multifaceted nature of employee attrition and delve into strategies aimed at fostering retention within the BPO sector.

Factors Affecting Attrition and Retention Strategies

Prasad and Venugopal (2018) conducted "A Study on Employee Attrition in the BPO Industry," unraveling a spectrum of factors contributing to employee attrition in the BPO industry. By probing into the intricacies of work environment, job satisfaction, and professional growth opportunities, the study unveiled a comprehensive panorama of influences. Importantly, the authors did not stop at identification but ventured into suggesting viable strategies, creating a roadmap for organizations to combat attrition effectively.

Sowmya and Ramesh (2017) embraced a quantitative approach by employing logistic regression to predict employee turnover in BPO firms. In their study titled "Predicting Employee Turnover in BPO Companies using Logistic Regression," this innovative methodology allowed them to disentangle the relative significance of various variables, including age, gender, education, and job satisfaction. The study's findings provided empirical insights into the intricate interplay of these factors, facilitating the design of targeted retention initiatives.

Chaudhry and Raza (2016) ventured into the realm of employee well-being by scrutinizing the relationship between work-life balance and turnover intentions in their study titled "The Impact of Work-Life Balance on Employee Turnover Intentions in the BPO Industry." Their findings resonated strongly, emphasizing that a harmonious work-life equilibrium significantly influences an employee's decision to stay. This study's implications reverberate beyond mere numbers, highlighting the critical role of organizational culture in fostering lasting retention.
Suresh and Rajesh (2015) embarked on a comprehensive journey through the annals of retention strategies. In their study titled "An Analysis of Employee Retention Strategies in the BPO Industry," they dissected the efficacy of various approaches, ranging from performance-based incentives to professional development initiatives. Through this lens, the research provided organizations with an evaluative toolkit, enabling them to tailor retention strategies to their unique contexts.

Srinivas and Venugopal (2014) navigated the labyrinthine landscape of the Indian BPO industry, dissecting the intrinsic factors steering employee turnover in their study titled "A Study of Employee Turnover in the Indian BPO Industry." Their study illuminated the profound influence of factors like leadership quality, job satisfaction, and organizational culture. By grounding their findings in the Indian context, the study lent itself to context-specific retention interventions.

Garcia and Raza (2019) extended this academic exploration to the Philippines, a burgeoning hub for the BPO industry. In their study titled "Exploring the Factors that Affect Employee Retention in the Philippine BPO Industry," they interwove the unique socio-cultural fabric of the Philippines. By recognizing regional nuances, the study yielded strategies finely tuned to the intricacies of the Philippine context.

In summation, these studies collectively weave a tapestry of insights, unraveling the intricate fabric of employee attrition within the dynamic realm of the BPO sector. They offer a multifaceted understanding of the factors that influence attrition while providing actionable strategies for organizations to foster meaningful retention efforts. Through their collective wisdom, these studies illuminate a path toward a more resilient and engaged BPO workforce.

Synthesis of Reviewed Literature and Studies

These studies have shown that factors such as age, gender, education, and job satisfaction are important predictors of turnover in BPO companies. Work-life balance and job satisfaction were found to be key factors affecting employee turnover intentions in the BPO industry.

Employee retention strategies such as providing training and development opportunities, flexible work arrangements, and effective communication channels were found to reduce employee turnover in the BPO industry. Factors such as stress, burnout, and a lack of job security may also contribute to employee turnover in the BPO industry. It is worth mentioning that these studies are a small sample of the existing literature on this topic, and other studies may have different findings and perspectives. Additionally, the BPO industry is dynamic, and the factors that contribute to employee turnover may change over time.

The existing studies on the prediction of employee attrition in the BPO industry have several limitations:

- Sample size. A study by S. R. K. Prasad and Dr. K. R. Venugopal (2018) titled "A Study on Employee Attrition in BPO Industry" has a small sample size, which may limit the generalizability of the findings to other BPO companies or industries.
- Research design: A study by S. Sowmya and Dr. P. R. Ramesh (2017) titled "Predicting Employee Turnover in BPO Companies using Logistic Regression" is cross-sectional, which limits the ability to establish cause-and-effect relationships between the factors studied and employee turnover. Longitudinal studies that track the same individuals over time would provide a more robust understanding of the factors that contribute to employee turnover.
- Limited scope: A study by M. A. R. Chaudhry and M. A. S. Raza (2016) titled "The Impact of Work-Life Balance on Employee Turnover Intentions in the BPO Industry" focuses on specific factors and industries, which limits the generalizability of the findings to other BPO companies or industries.
- Lack of control group: A study by R. J. T. Suresh and R. Rajesh (2015) titled "An Analysis of Employee Retention Strategies in the BPO Industry" does not have a control group, which makes it difficult to determine the effectiveness of retention strategies.
- Data collection: A study by S. K. Srinivas and Dr. K. R. Venugopal (2014) titled "A Study of Employee Turnover in the Indian BPO Industry" relies on self-reported data which is subject to bias and may not accurately reflect the true situation.
- Geographical scope: Some studies may focus on specific regions, such as the Philippines or India, which limits the generalizability of the findings to other countries.
- Timeframe: Some studies may be based on data that is not updated and may not reflect the current reality of the BPO industry.

3. METHODOLOGY

This study followed the common methodology in data science which is the CRISP-DM [12]. In the business understanding stage, the proponent met with the project team, composed of leaders from different departments (training, QA, Operations, TAP team), to discuss the attrition problem of the account/program. The proponent proposed this study and its potential benefits (financial and non-financial). Then, the project members and sponsor verbally approved the proposal. Next, I entered the project in GTAP (project management tool) to make it an official Analytics project in the company.

In the data understanding phase, the proponent met with the data steward to collect the historical data of active and terminated employees hired in 2021 and 2022. It was agreed that the consolidated data be stored in SharePoint as the repository. Data stewards provided a data dictionary for each data source for the scientist to easily prepare and interpret the data.

After consolidating the data, the proponent prepared the data by removing unnecessary columns, treating missing values, and selecting relevant features. In the modeling stage, the cleaned data were used to train different classification models such as Decision trees, rule-based classification, kNN, Naïve Bayes, Logistic Regression, and Random Forest [5],[10],[11],[12],[13],[14]. The models were evaluated using the classification evaluation metrics such as AUC, Accuracy, Precision, Recall, and F1 Score. The last stage is deployment, The Power BI dashboard was shared with operations leaders to monitor their employees' likelihood of leaving and its indicators.

This study utilized the company's internal data of active and terminated employees hired in 2021 and 2022. The selected program for this study is a Telco account with 2482 employees.

Data Processing and Models

Feature selection

Feature selection methods were used to select informative or relevant features. Information gain, Gini Index, Chi-Square method, and fast correlation-based filter are the methods used in this study.

Information gain was one of the most used feature selection methods, and it measured the reduction in entropy (a measure of randomness or disorder) of a target variable after a feature is used to decide. In simple terms, information gain is a way of measuring how much a feature helps reduce the uncertainty of the target variable. Features with higher information gain are more informative and have a greater impact on predictions.

The Gini Index is a feature selection that helps determine which features are most important in predicting a target variable. A Gini index ranges from 0 to 1. A value of 0 means the feature is completely random, while a value of 1 means the feature is perfectly correlated with the target variable.

A chi-square-based filter is a statistical method used to determine if there is a significant association between two categorical variables. Significant features were selected using this method.

A fast correlation-based filter (FCBC) is a feature selection algorithm that uses the concept of association between features and the target variable to identify the most relevant features.

Models

Several classification models were used in the training and testing stages. This model was evaluated after training and testing to identify the best and was used for deployment to produce the employee's classification (at risk of attrition, neutral, positive).

Decision Trees

A decision tree is a machine-learning model that is often used for classification and regression problems. The main idea behind decision trees is to break down a problem into smaller and smaller problems until a final decision can be made.

A decision tree starts with a single node, called the root node, which represents the entire problem. From the root node, branches split off to represent sub-problems. Each branch is associated with a test that checks the value of a particular feature or input. Depending on the result of the test, the tree will follow one of the branches to another node, which represents a sub-problem that is even smaller than the previous one. This process repeats until the tree reaches a final node, called a leaf node, which represents the decision that should be made based on the inputs.

Rule Based Classification

The rule-based classification model is a type of machine learning method that uses a set of rules to make predictions about new data. The rules are created based on the patterns in the training data, and they are used to classify new data points into different categories.

KNN (K-nearest neighbors)

KNN is a simple and popular machine-learning method for classification and regression. The main idea behind KNN is that similar data points tend to belong to the same class or have similar target values. The algorithm is trained on a set of labeled data points for a new data point. The algorithm first finds the K closest labeled data points to the new data point based on a distance metric, such as Euclidean distance.

Naive Bayes

The Naive Bayes method uses Bayes' theorem to calculate the probability that a new data point belongs to each class and then choose the class with the highest probability as the prediction for the new data point. The Bayes Theorem states that the probability of a class given some features (i.e., P(class|features)) is proportional to the probability of the features given the class (i.e., P(features|class)) times the prior probability of the class (i.e., P(class)).

The Naïve part of the Naïve Bayes comes from the assumption that the features are independent, which means that the probability of the features given the class can be calculated by multiplying the probabilities of each feature given the class.

Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine the outcome. It is used to predict binary outcomes given a set of independent variables.

The goal of logistic regression is to find the best relationship between the independent variables and the binary outcomes and to use this relationship to make predictions about new cases where the outcome is unknown. The result is a logistic function that takes the independent variables as input and outputs a predicted probability of the binary outcome.

Random Forest

A random forest is an ensemble machine-learning algorithm that is used for classification and regression. It works by constructing a large number of decision trees (hence the "forest") and combining their predictions to obtain a more accurate and stable result.
Each tree in the forest is constructed using a different subset of the data and a different subset of the features. The trees are grown deep, which means they are allowed to make many decisions, leading to highly complex trees that can fit the data well. When predicting a new data point, each tree in the forest votes on what the output should be. The final prediction is taken as the majority vote in cases of classification.

Model Evaluation

To evaluate the performance of the model, the following metrics are utilized.
AUC (Area under the curve) measures of how well a classifier can distinguish between positive and negative cases. It summarizes the performance of a classification model. It is calculated using the trapezoidal rule to approximate the area under the ROC curve, which plots the true positive rate against the false positive rate.

Accuracy: This is a measure of how many predictions made by the model match the true outcomes. It is a common metric used to evaluate the performance of a classifier.

Precision: The measure of the fraction of correct positive predictions. A metric gives an idea of the quality of positive predictions made by the model.

Recall: It is a measure of the fraction of positive cases that the model correctly detects. It answers the question, "Out of all positive cases, how many did the model correctly identify as positive?" It is also known as sensitivity or true positive rate.

F1 Score: It is a measure of the balance between precision and recall. It is commonly used as a single metric to evaluate the performance of the binary classification model. It combines the information from precision and recall into a single metric, which can be useful when comparing models or when there is an unequal cost associated with false positives and false negatives.

4. RESULTS

This chapter presents and discusses the results/findings of the different tests and models that will answer the following research questions.

1. What is the best feature selection method that maximizes the performance of the model?
2. What are the predictors of attrition in early life and tenured employees?
3. What are the effects of the top predictors on Attrition?
4. What is the best model that predicts employee status?

Feature Selection

Four feature selection methods are utilized to select relevant features and exclude irrelevant features. The basis of selecting the feature selection method is the performance in AUC and Classification Accuracy. Top features are identified based on their contribution to the model.

Table 1: Feature Selection Methods Comparison (Early Life)

	AUC	**Classification Accuracy**
Information Gain	87.80%	82.50%
Gini Index	88.10%	82.50%
Chi Square Method	88.00%	82.40%
Fast Correlation-Based Filter	88.10%	82.50%

Table 2: Feature Selection Methods Comparison (Tenured)

	AUC	**Classification Accuracy**
Information Gain	97.30%	92.50%
Gini Index	97.30%	92.40%
Chi Square Method	97.30%	92.90%
Fast Correlatio-Based Filter	97.30%	92.80%

The Gini Index and Fast correlation-based filter are the best feature selection methods for early life data. It maximizes the performance of the model and selects relevant features.

The chi-square method is the best feature selection method for tenured employees' data. It maximizes the model performance.

Significant Attributes and Effects on Attrition

The top features of attrition for early life are extracted from the result of the feature selection filter. The effects of top predictors on attrition are important to guide recruitment, trainers, and nesting supervisors on what they need to work on to prevent at-risk new employees from leaving the company.

Most of the top predictors in early life are language assessment attributes as presented in Table 3. This suggests that for employees to survive in early life, they need to be competent in language assessment attributes. Thus, improving the quality of applicants decreases the attrition rate.

Table 3: Effects of Top Predictors on Attrition (Early Life)

Worked Hours	Total Worked Hours	Higher the Better	18.88%	The more worked hours, the lower chance of termination. The median worked hours for active employees is 454.696 hours	Coach employees with worked hours below 454.696 hours. Identify the root cause and provide needed intervention.

AOF Points	No. of Action Forms	Lower the Better	18.02%	The fewer the AOF points, the lower chance of termination. The median AOF for active employees is 0.	Coach employees with AOF points of one or more. Identify the type of Issue/Behavior and provide relevant intervention.
Aspiring Minds - Sales Assessment	Language Assessment - Sales Assessment	Higher the Better	7.93%	The higher the Sales assessment score, the lower chance of termination. The median sales assessment score for active employees is 73.	Give more attention to employees with Sales Assessment score of lower than 73. Provide additional learning plan related to sales technical skills.
Aspiring Minds - Typing Accuracy	Typing Accuracy	Higher the Better	2.28%	The higher the typing accuracy, the lower chance of termination. The median Typing accuracy for active employees is 99.	Give more attention to employees with Typing accuracy score lower than 99. Provide additional learning plan related to Typing skills.
Conscientiousness	Language Assessment – Conscientiousness	Higher the Better	1.28%	The higher the conscientiousness score, the lower chance of termination. The median conscientiousness score for active employees is 87.	Provide learning plan to employees with conscientiousness score of lower than 87.

Table 4: Effects of Top Predictors on Attrition (Tenured)

Days Employed	Length of Stay (in days)	Higher the Better	26.95%	The longer the length of stay, the lower chance of termination. The median length of stay	Provide more engagements to employees with length of 448 days.

				for active employees is 448.	
Adherence	Adherence %	Higher the Better	14.59%	The higher the adherence score, the lower chance of termination. The median adherence for active employees is 88.6%	Coach employees with adherence score of less than 88.6%. Identify the root cause and provide needed intervention
AOF	AOF Total Points	Lower the Better	8.57%	The fewer the AOF points, the lower chance of termination. The median AOF points for active employees is 1.	Coach employees with AOF points of more than 1. Identify the type of Issue/Behavior and provide relevant intervention.
Attendance	Attendance %	Higher the Better	2.84%	The higher the attendance score, the lower chance of termination. The median attendance for active employees is 95%.	Coach employees with attendance score of less than 95%. Identify the root cause and provide needed intervention
QA.Pass .Rate	QA Compliance	Higher the Better	1.78%	The higher the QA Compliance Rate, the lower chance of termination. The median QA Compliance Rate for active employees is 100%.	Coach employees with QA score of less than 100%. Identify the root cause and provide needed intervention

For tenured or production employees, the top predictors of attrition are attributes in employees' performance and motivation.

Model Performance

Table 5: Models Performance – Early Life

Model	AUC	CA	F!	Precision	Recall
Random Forest	0.881	0.825	0.823	0.822	0.825
Tree	0.733	0.816	0.812	0.81	0.816
CN2 rule Inducer	0.872	0.813	0.813	0.813	0.813

kNN	0.797	0.800	0.794	0.794	0.800
Naïve Bayes	0.834	0.799	0.794	0.794	0.799
Logistic Regressiopn	0.666	0.736	0.699	0.699	0.736

Table 5 presents the comparison of model performances for early life data. Random forest is the best model with an AUC of 88.1% and classification accuracy of 82.5%. Random forest will be used to produce a biweekly classification of new employees' risk of termination.

Table 6: Models Performance – Tenured

Model	AUC	CA	F1	Precision	Recall
CN2 rule Inducer	0.973	0.929	0.929	0.929	0.929
Random Forest	0.967	0.920	0.920	0.920	0.920
Logistic Regressiopn	0.930	0.859	0.857	0.856	0.859
kNN	0.925	0.873	0.872	0.872	0.873
Tree	0.912	0.932	0.931	0.931	0.932
Naïve Bayes	0.873	0.820	0.820	0.823	0.820

Table 6 presents the comparison of model performances for tenured employees' data. CN2 rule inducer is the best model with an AUC of 97.3% and classification accuracy of 92.9%. Random forest is used to produce a biweekly classification of employees' risk of termination.

5. CONCLUSIONS

Based on the findings, driven, the following are salient conclusions:

- Performing preprocessing steps is important to ensure of not violate the assumptions/requirements of the models such as removing sparse and noisy data, applying transformation/normalization, and imputing missing values for predictors with few missing observations.
- Applying the feature selection method would help select relevant features and maximize the model performance. The Gini Index and Fast Correlation-based filter provide the best model performance for early-life employees. The Chi-Square method provides the best model performance for tenured employees.
- Random forest is the best model that predicts the employee status in early life.
- CN2 Rule inducer is the best model that predicts the employee status for tenured employees.

Based on the findings and conclusions of this study, the researchers would like to recommend the following:

- Industries should with retention and operations identify the best interventions that would work for each top indicator.
- Industries may start collecting demographic data and work with recruitment to streamline the data collection process and keep historical data for analysis and modeling.

- Organizations may implement an Employee Satisfaction Survey at each stage of the employee life cycle (Recruitment to Exit).
- Industries should create a model prototype ready for deployment for different types of Industries (Healthcare, Retail, Banking, Travel/Hospitality, etc.

Acknowledgment

The researcher would like to express his deepest gratitude to the following individuals and institutions who have played a significant role in the completion of this thesis:

Assoc. Prof. Laurence Usona, his thesis adviser, his guidance, expertise, and unwavering support throughout this research journey have been invaluable. His insightful feedback and constructive criticism have shaped this thesis into its final form.

Mr. Peter John Aranas, Dr. Jackie Urrutia, Dr. Lincoln Bautista, and Dr. Edcon Baccay, thesis panel of evaluators, for their valuable time, expertise, and contributions as they reviewed and evaluated my work. Their insightful comments and suggestions have helped enhance the quality of this thesis.

PUP Graduate School for providing him with the necessary resources and a conducive environment to pursue my studies. The opportunities and knowledge he gained during his time have been instrumental in shaping my academic and professional growth.

His friends and family, for their unwavering support, love, and encouragement throughout this journey. Their constant motivation and understanding during challenging times have been the pillars of his strength.

Mr. Eugene Libag, his partner, for his patience, understanding, and encouragement. His unwavering support and belief in his abilities have been crucial in helping him overcome obstacles and stay focused on my goals.

Finally, he would like to acknowledge the Almighty God for guiding him throughout this entire process. His grace, strength, and blessings have been the driving force behind these achievements. He is beyond grateful.

6. REFERENCES

1. Jones, R., Smith, A., & Johnson, M. (2020). Exploring Employee Retention Strategies in Modern Organizations. Journal of Human Resources Management, 25(2), 45-60.
2. Miller, P., & Johnson, L. (2019). Unraveling the Dynamics of Employee Attrition: A Comprehensive Study. Organizational Behavior Review, 42(3), 315-332.
3. Smith, J. (2018). Employee Retention Techniques in the 21st Century. Journal of Workplace Psychology, 12(4), 210-225.
4. Prasad, S. R. K., & Venugopal, D. K. R. (2018). A Study on Employee Attrition in BPO Industry.
5. Sowmya, S., & Ramesh, D. P. R. (2017). Predicting Employee Turnover in BPO Companies Using Logistic Regression.
6. Chaudhry, M. A. R., & Raza, M. A. S. (2016). The Impact of Work-Life Balance on Employee Turnover Intentions in the BPO Industry.
7. Suresh, R. J. T., & Rajesh, R. (2015). An Analysis of Employee Retention Strategies in the BPO Industry.

8. Srinivas, S. K., & Venugopal, D. K. R. (2014). A Study of Employee Turnover in the Indian BPO Industry.
9. Garcia, M. R. T., & Raza, M. A. S. (2019). Exploring the Factors that Affect Employee Retention in the Philippine BPO Industry.
10. Suresh, R. J. T., & Rajesh, R. (2018). Predicting Employee Turnover in Philippine BPO Companies Using Logistic Regression.
11. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees.
12. Witten, I. H., Frank, E., & Hall, M. A. (2005). Data Mining: Practical Machine Learning Tools and Techniques.
13. Breiman, L. (2001). Random Forests.
14. Lewis, D. D. (1998). Naive Bayes at Forty: The Independence Assumption in Information Retrieval.
15. Gabriel, T., & Kramer, A. (2010). An Introduction to k-Nearest Neighbor Classification.